Robust data management for “big clinical data”

Dave Kale

CSCI 578, March 22, 2011
Similar patients, different care?

Fred

Admitted to CHLA PICU

George

Admitted to different ICU

Experienced peds intensivist

Limited relevant experience

3/24/11
Similar patients, different care?

Fred
Experienced peds intensivist
Admitted to CHLA PICU

George
Limited relevant experience
Admitted to different ICU
Similar patients, different care?

**Reality:** quality of care varies across institutions and individual providers*

Data-driven decision support

Experienced peds intensivist

Limited relevant experience
Who am I?

David Kale

BS in Symbolic Systems, Stanford
MS in Computer Science (AI), Stanford

• Applied machine learning guy at CHLA
• Committer on Apache OODT project
• Here to talk about software architecture for management of “big” clinical data
• Not a software architecture expert
  (learning as I go)
The Whittier VPICU

- **Laura P. and Leland K. Whittier Virtual Pediatric Intensive Care Unit (VPICU)**
  - Founded in 1998 by Randall Wetzel at CHLA
  - Funded by donation from Whittier foundation, grants
  - Multidisciplinary team (clinicians, clinical researchers, statisticians, computer scientists, etc.)

**Mission:**

*Improve the care of critically ill children by leveraging information technology* to drive innovative understanding of critical illness and to share this understanding with critical care providers (via telemedicine, distance learning, medical informatics, data mining, novel decision support tools, etc.).
Short detour into polemics: the EHR!
Electronic health records (EHRs)

As defined by Department of Health and Human Services, meaningful use of digital health:

• Accurate and complete information about a patient's health.
• The ability to better coordinate the care they give.
• A way to securely share information with patients and their family caregivers over the Internet, for patients who opt for this convenience.
• Information to help doctors diagnose health problems sooner, reduce medical errors, and provide safer care at lower costs.
Electronic health records (EHRs)

As defined by Department of Health and Human Services, meaningful use of digital health:

• Accurate and complete information about a patient's health.
• The ability to better coordinate the care they give.
• A way to securely share information with patients and their family caregivers over the Internet, for patients who opt for this convenience.
• Information to help doctors diagnose health problems sooner, reduce medical errors, and provide safer care at lower costs.

3/24/11

http://healthit.hhs.gov/portal/server.pt/community/healthit_hhs_gov__home/1204
Don’t believe the hype...

As defined by Department of Health and Human Services, meaningful use of digital health:

• **Don’t improve quality of care:** Romano and Stafford, Arch Intern Med, Jan 2011

• **Mixed results on quality improvement:** Jones, et al., Am J Manage Care, Dec 2010

• **Large gap between postulated and demonstrated results:** Black, et al., PLoS Medicine, Jan 2011

• **No association between EHR and quality:** Zhou, et al., J AMIA, July/August 2009
The “moral” of the detour

- Any EHR (product) **better** than **no** EHR (product)?
  - Concept/idea vs. reality/implementation

- At least we’re collecting lots of data now, right?
  - worth millions/billions spent on them?
  - good public policy?

- EHRs should and **can**
  - improve quality of care
  - improve efficiency
  - reduce cost
  - transform the practice of medicine
Transforming care requires

1. Large stores of highly detailed, granular data
2. Robust, flexible data management
3. Algorithms to turn data into information
4. Intuitive delivery and display of that information
5. Clinicians and researchers who can use these tools

March 24, 2011
Today we will discuss

1. Large stores of highly detailed, granular data

2. Robust, flexible data management

3. Algorithms to turn data into information

4. Intuitive delivery and display of that information

5. Clinicians and researchers who can use these tools
(3.)-(5.) in a nutshell

3. Algorithms to turn data into information
   - Failure to adopt methods from modern machine learning, tools from “big data” analytics (e.g., Hadoop)
   - Most “decision support” is “rules” (not data) driven
   - Sadly, ignorant of (or resistant to) F.O.S.S.

4. Intuitive delivery and display of that information
   - Do these companies hire UI designers???
   - Simply regurgitate data or spit out simple reports

5. Clinicians and researchers who can use these tools
   - Missing/incomplete training in computers, statistics, etc.
   - Negative feedback loop between docs, “tool” designers

March 24, 2011
Large stores of digital data

• **EHRs increasingly good at this!**
  – Medications ordered and administered (and when)
  – Notes, observations, outcomes, diagnoses, etc.
  – Physiological data, sometimes even waveforms

• **Not good at getting data out for analysis**
  – Proprietary, inaccessible, highly restricted, “vendor lock”
  – Serve “niche” purpose
  – Valuable data haphazardly fragmented across stores

• **Most current EHRs are**
  – Electronic version of the old paper record
  – Complicated, rudimentary database “view”
Example: Hospital-wide EHR

Hospital

EHR Application

EHR Database
Example: Hospital-wide EHR

- Unnerving but many legitimate reasons for this setup
- Works well for small transactions
- No major disruptions to service (that I’ve ever heard of)
- Even provide “amenities”
  - Reporting stuff on top
  - Will sell us a data warehousing product
  - Direct backend connection, if we want
Example: Hospital-wide EHR

- **Vendor lock**
  - Must learn proprietary query language
  - Warehouse product costs $$$$$$$$$$
  - What about non-canned analysis?
- **Hospital also erects barriers (legitimate but annoying)**
  - Concerns about privacy, data security, etc.
  - Worries about production systems
- **What about data from other sources?**

---

**Wasn’t designed for our use case:**
- Large-scale, experimental analysis
- “DIY” research efforts

**Vendor, hospital surprised we want the data...???”
Data sources at CHLA

- EHR (remote RDBMS)
- Homegrown system from 80s
- Files with proprietary compression
- MS Access extracts

March 24, 2011
Data sources at CHLA

- EHR (remote RDBMS)
- Homegrown system from 80s
- Files with proprietary compression
- Historical data (RDBMS)
- MS Access extracts
- Excel spreadsheets
- Census data
Data sources at CHLA

- EHR (remote RDBMS)
- Homegrown system from 80s
- Files with proprietary compression
- Historical data (RDBMS)
- Excel spreadsheets
- Census data

MS Access extracts
We have a lot of data...

...stored in a lot of very different places
...using a lot of different technologies
...described by a lot of different schema
...and still growing!

- “Vertically,” “horizontally” partitioned (haphazardly)
- Can almost never use it in place

- Sounds like ETL into a data warehouse
- Not quite: must serve diverse, ad hoc research whose needs cannot always be anticipated ahead of time

March 24, 2011
Dr. Smith wants to see whether shock index (SI) within the first six hours of admission to PICU is predictive of length of stay or mortality.  $SI = \frac{HR}{SBP}$

- Admission, HR and SBP data in different databases
- Finding data within window requires extra work
- SI assumes data is complete, time aligned
- Probably should pre-compute, store SI somewhere
- Perhaps deploy a research database storing common metrics, scores, etc., for all patient episodes
**Example research question #2**

Dr. Jones wants to know whether treating patients with drug X before an operation helps prevent occurrence of bradycardia after anesthesia, intubation.

- Finding bradycardias (docs “know it when they see it”)
- High frequency physiological time series data stored in thousands of individual files
- Candidate for distributed map/reduce (Hadoop)
- Bradycardia discovery needs to be iterative, interactive
  - Train initial classifier or ranking function, return results
  - Let doctor evaluate results, then re-train

March 24, 2011
Dr. Adams wants to examine health disparities based on where patients live and other demographics.

- Outcomes, addresses in separate databases
- Data then has to be geocoded
- Then US census data mapped to geocoded data
- Should automate geocoding of new data as it arrives
- Geospatial search engine / database?
Example research question #4

Dave wants to find the K nearest neighbors to a “query” PICU episode (multivariate time series of physiological variables plus some demographics).

- Can’t compare as is; must extract features for compare
- Framework allows user to adjust parameters?
  - Compare limited time windows (can’t pre-compute features)
  - Try different distance metrics (can’t pre-compute distances)
- Perhaps want to learn (“mine”) good features?
- Another good candidate for map/reduce (Mahout)
- “Holy grail” problem (architecture + machine learning)

March 24, 2011
Diverse research needs

• Often different data set for each question
• Analysis cannot be performed \textit{in situ} (in database)
  – Most clinical researchers want own copy of data
  – Programmers (like me) want to write their own
• Often match / integrate / combine data from \textgreater 1 source
• Queries rarely simple or straightforward
• Data often must be transformed so \textit{provenance is paramount}

• Not a simple “ETL into denormalized warehouse”
• \textit{Rapidly deploy, maintain multiple “warehouses”}
Robust data management?
Robust data management

We favor a configurable software-based architecture ("deconstructed ETL")

March 24, 2011
Extract from original sources

- Periodic extractions to “staging” files
  - Files are universal data connectors
  - Stored on our hardware
  - Minimal transformation; just get data
  - Schedule to minimize impact

- Stand up OODT product service, grid components in front of data sources
  - Extraction via `wget`, `curl`, etc.
  - Can be re-used by other tools!

- Protected health information (PHI) so must be highly secure
“Heterogeneous data products” archive

- OODT CAS File Manager
  - ingests, organizes products
  - generates metadata
  - builds catalogue
  - allows browsing of ingested products

- Deploy “workflows”
  - Verify data, detect anomalies
  - Integrate “cross-source” data
  - Generate “downstream” products
  - Map data to ontology

- Still PHI so must be highly secure
Ingest HDPs into research DBs (RDBs)
- specs determined by research
- targeted vs. re-useable
- most described by ontology

Built using workflows
- New RDB may require manual effort
- Try to build workflows from re-useable pieces

One part of RDB workflow may be de-identification of data, so that less security/access control required
“Deconstructed ETL”

Web services allow variety of ways to get at RDBs
- APIs
- Web front-ends
Distributed data-sharing network

Other sites have similar (not necessarily identical) architectures

Query Services

APIs for analytical software/routines (R, MATLAB, etc.)

Users

Web Interface for querying, downloading data

Public Front End

Public Back End

Ingest

Research data warehouses

Common Schema (based on ontology)

Data served by product services + web services

OODT Data Grid Srvcs

CAS Filemgr

Research DB1

CAS Filemgr

File Catalog

CAS Filemgr

Research DB2

Other VPICU Sites
Overall architecture

CHLA/JPL VPICU Research Data Repository Architecture

Other sites have similar (not necessarily identical) architectures

CHLA

Original/Intermediate Schema

Common Schema (based on ontology)

Research data warehouses

Data served by product services + web services

OODT Data Grid Srvcs

Query Services

APIs for analytical software/routines (R, MATLAB, etc.)

Web Interface for querying, downloading data

Users

Ingest

EHR

Custom DBs

Clinical Apps

Raw Monitor Data

Non-VPICU Hardware/Apps

Extraction via product services + web services (XMLPS, grid)

(Distributed?) File system

"Warehoused" Heterogeneous Data Products

Managed by CAS FileManager, etc.

CHLA Data Sources

Ingest

Internal Back End

Public Back End

Other VPICU Sites

3/24/11
Transforming care requires

1. Large stores of highly detailed, granular data
   - With EHRs, data is flooding in, piling up
   - Proprietary, fragmented, hard to get/integrate

2. Robust, flexible data management
   - Must serve many ad hoc research needs we can’t anticipate
   - Canned, all-in-one ETL solutions won’t work
   - Must be highly configurable, composed of re-useable pieces

Prerequisite for the other three parts:
   - Algorithms to turn data into information
   - Intuitive delivery and display of that information
   - Clinicians and researchers who can use these tools

Ongoing work, very high impact, lot to be done. We are at the cutting edge.

March 24, 2011
Exciting things going on in VPICU!

• NLM-sponsored symposium *Meaningful Use of Complex Medical Data*
  – August 26-28, at CHLA, schedule TBA
  – Website under construction: [http://www.mucmd.org](http://www.mucmd.org)
  – Call for abstracts and papers coming
  – Speakers will include experts from *medicine, software architecture, artificial intelligence, machine learning, data mining, open source software*

• **VPICU is hiring! For both full-time and internships!**
  – Send inquiries and resumes to [dkale@chla.usc.edu](mailto:dkale@chla.usc.edu)
  – Look for announcement on USC job lists